ALIGNMENT, VALIDITY, AND RELIABILITY OF THE SPRING 2000 GOLDEN STATE EXAMINATIONS

A REPORT TO THE
SENATE
ASSEMBLY
DEPARMENT OF FINANCE
STATE BOARD OF EDUCATION

PREPARED BY THE CALIFORNIA DEPARTMENT OF EDUCATION
STANDARDS AND ASSESSMENT DIVISION
APRIL 13, 2000

Alignment, Validity, and Reliability of the Spring 2000 Golden State Examinations

Executive Summary

In 1999, the California Legislature allocated funding to ensure that the spring 2000 Golden State Examinations (GSE) were aligned to California's content standards and that the examinations were valid and met industry standards for reliability. By early spring of 2000 a series of meetings to ensure GSE alignment had been conducted, as well as reviews of the validity and analyses of reliability of the spring 2000 GSE.

Alignment to Standards

Independent review panels evaluated the alignment of items to content standards. The panels looked for direct, explicit relationships between items and standards. To verify alignment, they required that the knowledge, information, activity, and skill described in each standard be addressed by the item. The panels concluded that the spring 2000 Golden State Examinations were aligned to the content standards and that these examinations were appropriate for administration in spring 2000. These findings confirmed the conclusions of the GSE development team coordinators and team leaders earlier in the month. The review panels further found that all examinations in history/social science, mathematics, and science substantially covered the breadth of the standards, with high quality, well-written items, and that these examinations would appropriately measure the student's mastery of the content standards.

The independent review panels concluded that two examinations administered in winter 2000, reading/literature and written composition, were not yet fully aligned. Since these examinations are administered only in winter, the next administration will be in winter of 2001. The independent reviewers' recommendations will provide the GSE development teams with excellent guidance in the development and field testing of new items and will ensure that for the next administration all test items on these examinations will be fully aligned to standards.

Validity

The independent review panels evaluated the content validity of the winter and spring 2000 Golden State Examinations by examining all test items, assessing their alignment to standards, and considering whether the items covered an appropriate range of standards. If the panels were able to identify direct relationships between items and standards, confirm that the knowledge, information, activity, and skill described in each standard was addressed by the item, and verify that the test as a whole demonstrated appropriate standards coverage, the test was judged to have content validity.

The panels found that the spring 2000 Golden State Examinations were aligned to standards and that the tests demonstrated appropriate standards coverage. They recommended that in the future, coverage of certain content strands in history/social science, mathematics, and science be varied to ensure coverage of the standards over time. They recommended that both alignment and standards coverage need to be improved on the reading/literature and written composition exams.

The California Department of Education (CDE) plans to use the panel recommendations to establish the validity of the reading/literature and written composition exams to be administered in winter 2001 and to ensure the continuing validity of the Golden State Examinations in the other content areas.

Reliability

The GSE review process included a technical review designed to indicate whether the Golden State Examinations met industry standards for reliability. Although reliability data for the spring 2000 exams are not yet available, data for spring 1999 indicate that the Golden State Examinations provide accurate scores at the level critical for identifying students who qualify for honors recognition. These data also indicate the reliabilities of several Golden State Examinations need to be improved, particularly if the exams will be used for high stakes decision-making.

A number of ongoing initiatives to increase reliabilities are already being implemented. These include increasing the number of written-responses required of students on certain exams from one to two, replacing holistic with component scoring on written-response and lab tasks, and converting students' multiple-choice and written-response scores to a common scale.

The technical analyses suggest that several additional measures could be undertaken to increase GSE reliabilities. These include increasing the number of multiple-choice items on the examinations, which may entail an increase in testing time. These measures would require modifications in GSE test designs but would lead to greater accuracy in the test scores. This may be an opportune time to implement such revisions in the Golden State Examinations.

Alignment, Validity, and Reliability of the Spring 2000 Golden State Examinations

Introduction:

Assembly Bill 265, signed into law in October 1995, provided for the development of new statewide content and performance standards in the core curriculum areas of language arts, mathematics, history/social science, and science. The State Board of Education (SBE) has adopted new statewide content standards, although performance standards are not yet developed. In 1999, Senate Bill 160 provided funding to support alignment of the Golden State Examination (GSE) to the new standards and to ensure that the examinations are valid and meet industry standards for reliability. In fall of 1999, a process was implemented to ensure that the Golden State Examinations are aligned to standards, reliable to the level of industry technical standards, and valid. This report summarizes that process and the progress made toward meeting this goal.

Background:

The GSE Program was established in 1983 by Senate Bill 813/Chapter 498 to offer rigorous examinations in key academic subjects to students in grades 7-12 and to recognize students who demonstrate outstanding achievement on each examination. GSE was reauthorized in 1991 by Senate Bill 662/Chapter 760, and reenacted in 1995 by Assembly Bill 265/Chapter 950. The GSE recognizes students who achieve high honors, honors, and recognition levels of achievement on each examination. Students who meet or surpass these levels are recognized as Golden State Scholars. All Golden State Scholars receive academic excellence awards from the state, and high honors and honors designees receive a gold insignia on their diplomas. Notice of success on the GSE becomes part of a student’s permanent transcript, signifying high achievement to colleges, universities, and employers.

In 1996, Assembly Bill 3488 established the Golden State Seal Merit Diploma to recognize graduates who have mastered the high school curriculum. To qualify for this diploma, a student must achieve high honors, honors, or recognition on at least six Golden State Examinations. These examinations must include written composition or reading/literature, U.S. history, a mathematics exam, a science exam, and two other examinations of the student’s choice. More than 1,415 Golden State Diplomas were awarded to qualifying 1997 graduates, and 2,722 were awarded in 1998. More than 4,720 diplomas have been awarded to 1999 graduates.

The first Golden State Examinations, first-year algebra and geometry, were offered in 1987; examinations are also now offered in U.S. history, economics, biology, chemistry, second-year coordinated science, written composition, government/civics, reading/literature, high school mathematics, physics, and Spanish language. The examinations assess students’ knowledge of the subjects and their application of that knowledge. Examinations include multiple-choice and written-response questions. The science examinations also include lab tasks. Further information on the Golden State Examinations is available in the GSE Teacher Guides, question and answer documents, and other materials attached at the end of this report.

Until 1999, the Golden State Examinations were aligned to content-area frameworks for California public schools and standards developed by the Education Roundtable. The appropriation provided in Senate Bill 160, however, funded a standards-alignment initiative to bring the spring 2000 Golden State Examinations into alignment with statewide content standards. Processes and results of this initiative are described in the following pages.

The discussion that follows is divided into four major sections.

- Section One - describes the processes used to align the winter and spring 2000 Golden State Examinations to the content standards and results of that alignment effort.
- Section Two - discusses validity as it applies to the Golden State Examinations.
- Section Three - discusses the reliabilities of the Golden State Examinations.
- Section Four - provides the conclusion.

Section One—Alignment Process

Alignment of the spring 2000 Golden State Examinations began in Sacramento in fall 1999 when panels of members with content-area expertise were convened to examine the extent to which GSE items were aligned to the statewide content standards and develop recommendations for improving alignment. Guided by these recommendations, GSE development teams selected items for the spring 2000 examinations. In early March 2000, GSE development team coordinators and team leaders met in Sacramento to reexamine spring 2000 exam alignment efforts. Later that month, the exam alignments underwent further evaluation by panels of independent reviewers. Lists of participants in the fall 1999 and spring 2000 independent review meetings are presented in Appendix A.

The processes used by panelists at each meeting to determine alignment were nearly identical. They first reviewed the standards for their content areas. Then they examined the test items and evaluated their alignment to the standards. Their evaluations took into account how fully the content of an item matched the content of a standard, how well the cognitive activity or skill required by the item matched the cognitive activity or skill required by the standard, and whether the items as a whole addressed a range of standards.

Alignment Results for Winter and Spring 2000 Golden State Examinations

In March, the independent review panels concluded that the spring 2000 Golden State Examinations were aligned to the content standards and that these examinations were appropriate for administration in May 2000. These findings confirmed the conclusions made by the GSE development team coordinators and team leaders earlier in the month. In fact, the independent panels often found relationships to standards in addition to those identified by the development teams. The independent panels further found that all examinations in history/social science, mathematics, and science substantially covered the breadth of the standards with high quality, well-written items and could measure the student's extent of mastery of the content standards.

In addition, the independent review panels determined that two examinations administered in winter 2000, reading/literature and written composition, were not yet fully aligned. Since these examinations are administered only in winter, the next administration will be in winter of 2001. Guided by the independent reviewers' recommendations, GSE item development teams will be able to develop and field test new items and ensure that for the next administration items are also fully aligned to standards.

Each independent review panel was encouraged to offer recommendations for enhancing the GSE that they reviewed. They offered the following suggestions for improving future examinations:

History/Social Science

U.S. History Examination. The panel found that all items were aligned to the standards, and they recommended that more items address elements of minority history, that items should be developed for standard 11.3: The role religion played in the founding of America, and that fewer items address standard 11.1: Events surrounding the founding of the nation.

Government/Civics Examination. The panel members found that all items were aligned to the standards, and they recommended that more questions address standards 12.1: Principles and moral values of American democracy, and 12.8: The influence of the media on American political life.

Language Arts

Reading/Literature Examination. The panel members found that the winter 2000 GSE in Reading/Literature was fairly well aligned to standards. They recommended for the winter 2001 administration, however, that both multiple-choice and written-response items cover a broader range of standards in all three reading strands and that reading materials include nonfiction informational reading, poetry, and drama. They also recommended that the written-response prompts more clearly reflect the standards.

Written Composition Examination. The panel found that the winter 2000 GSE in Written Composition was partially aligned. Although all items aligned with standards, the panel thought that the items overall aligned with standards at lower grade levels than would be recommended. For the winter 2001 exam, they recommended that both multiple-choice and written-response items cover more standards in writing strategies and conventions and that a fuller array of grades 9 and 10 standards in writing conventions be tested. They also recommended that over time all writing application standards be tested and that writing prompts more clearly reflect the standards and provide clearer directions to the student.

Mathematics

First-year Algebra Examination. The panel found that all items were aligned to the standards and suggested that the number of standards addressed in geometry and in probability and statistics be reduced to allow maximum coverage of first-year algebra standards.

Geometry Examination. The panel found that all items were aligned to the standards and that the examination would be enhanced if students had opportunities to write proofs and to use trigonometric functions to solve for an unknown length of a side of a triangle.

High School Mathematics Examination. The panel found that all items were aligned to the standards and suggested that there be a better balance among items addressing first-year algebra, geometry, algebra II, and probability and statistics.

Science

Biology Examination. While all items were aligned with standards, the panel recommended that certain important areas of the standards be covered more thoroughly.

Chemistry Examination. The panel found that all items aligned with standards and that two strands of organic chemistry and nuclear chemistry were not addressed. They noted that the laboratory task is a complex activity that requires a broad knowledge of chemistry.

Coordinated Science Examination. The GSE in Coordinated Science reflects the courses approved for UC admissions. At present, there is no SBE policy about which standards students must meet for coordinated science. The panel found all items to be related to standards with about two-thirds of the standards addressed by items. They found two items, however, that aligned best to sixth-grade standards with one of these marginally aligned to chemistry and physics standards. Overall, the panel found this to be a fair test of the coordinated science courses as they are presently structured.

Physics Examination. The panel found that items were well distributed among the standards. They found that each strand was addressed but that not all standards within a strand were directly addressed. They found the lab task aligned to several content and investigation and experimentation standards and noted that the latter were addressed in ways that could not be addressed well by multiple-choice items.

Alignment Tables

To illustrate winter and spring 2000 GSE alignment, GSE development team coordinators produced content maps showing standards coverage. Summary versions of these maps are shown in the tables below. They include in the left column the strands in each content area and in the right column the percentage of multiple-choice items from the winter and spring 2000 exams that addresses the standards in that strand. Note that some items have been counted more than once because they address more than one standard.

Where a written-response task also addresses standards in a strand, this is noted. A table for the GSE in Spanish is included here, although this exam has been aligned to national rather than to California standards. California standards for Spanish have not yet been developed.

History/Social Science

The Golden State Examinations in history/social science include U.S. history, government/civics, and economics. Economics and government/civics examinations were administered in winter 2000. The U.S. history examination will be administered in spring 2000. The strand statements in the economics, government/civics, and U.S. history tables below have been shortened for ease of presentation.

Economics

In addition to the strands that address content, the economics standards include one of the strands of history and social science analysis skills shown in the table below. Since the analysis skills are assessed through content questions, percentages of multiple-choice items have not been calculated for these strands. As the table indicates, the GSE in Economics contains items that address the historical interpretation analysis strand.

Economics Content Strand	Percentage of multiple-choice items per content strand
Principles of Economics 12.1.0 Economic terms and concepts and economic reasoning	12%
12.2.0 Elements of the United States market economy in a global setting	35% (strand also assessed by written response)
12.3.0 The influence of the U.S. government on the American economy	25% (strand also assessed by written response)
12.4.0 Elements of the U.S. labor market in a global setting	6%
12.5.0 The aggregate economic behavior of the United States	14%
12.6.0 Issues of international trade, including how the U.S. economy affects, and is affected by, economic forces beyond its borders	8%
Total	100%
Historical and Social Sciences Analysis Skills • Chronological and Spatial Thinking	
• Historical Research, Evidence, and Point of View	
• Historical Interpretation	*

*As the SBE requires, history and social science analysis skills are assessed in connection with and through content questions.

Government/Civics

In addition to the strands that address content, the government/civics standards include the three strands of history and social science analysis skills shown in the table below. These analysis skills are assessed through content questions, so that percentages of multiple-choice items have not been calculated for these strands.

Government/Civics Content Strand	Percentage of multiple-choice items per content strand
Principles of American Democracy	
12.1.0 Principles and moral values of American democracy	6% (strand also assessed by written response)
12.2.0 The scope and limits of rights and obligations as democratic citizens	10% (strand also assessed by written response)
12.3.0 The fundamental values and principles of civil society	This strand is assessed by written response.
12.4.0 The unique roles and responsibilities of the three branches of government	28% (strand also assessed by written response)
12.5.0 Landmark U.S. Supreme Court interpretations of the U.S. Constitution and its amendments	10% (strand also assessed by written response)
12.6.0 Issues regarding campaigns for national, state, and local elective office	18% (strand also assessed by written response)
12.7.0 Powers and procedures of the national, state, tribal, and local governments	16%
12.8.0 The influence of the media on American political life	
12.9.0 Origins, characteristics, and development of different political systems across time	10%
12.10.0 Tensions within the U.S. constitutional democracy	2%
Total	100%
Historical and Social Sciences Analysis Skills	
• Chronological and Spatial Thinking	*
• Historical Research, Evidence, and Point of View	*
• Historical Interpretation	*

*As the SBE requires, history and social science analysis skills are assessed in connection with and through content questions.

U.S. History

In addition to the strands that address content, the U.S. history standards include the three strands of history and social science analysis skills shown in the table below. These analysis skills are assessed through content questions, so that percentages of multiple-choice items have not been calculated for these strands.

U.S. History Content Strand	Percentage of multiple-choice items per content strand
United States History and Geography: Continuity and Change in the Twentieth Century 11.1.0 Events surrounding the founding of the nation	14% (strand also assessed by written response)
11.2.0 Immigration from Southern and Eastern Europe	13% (strand also assessed by written response)
11.3.0 The role religion played in the founding of America	
11.4.0 The rise of the U.S. as a 20th century world power	14% (strand also assessed by written response)
11.5.0 The major political, social, economic, technological, and cultural developments of the 1920s	9%
11.6.0 The Great Depression and the role of the New Deal	9%
11.7.0 American participation in World War II	6%
11.8.0 The economic boom and social transformation of post-World War II America	12% (strand also assessed by written response)
11.9.0 United States foreign policy since World War II	12% (strand also assessed by written response)
11.10.0 Federal civil rights and voting rights developments	8%
11.11.0 The major social problems and domestic policy issues in contemporary American society	3%
Total	100%
Historical and Social Sciences Analysis Skills • Chronological and Spatial Thinking	*
• Historical Research, Evidence, and Point of View	*
• Historical Interpretation	*

*As the SBE requires, history and social science analysis skills are assessed in connection with and through content questions.

Language Arts

Reading/Literature

The GSE in Reading/Literature assesses the reading/literature standards for grades 9-10 and 11-12 in the percentages shown in the table below. This test was administered in winter 2000.

Reading/Literature Content Strand	Percentage of multiple-choice items per content strand
Reading (Grades Nine and Ten) 1.0 Word Analysis, Fluency, and Systematic Vocabulary Development	24%
2.0 Reading Comprehension	10% (strand also assessed by written-response)
3.0 Literary Response and Analysis	39% (strand also assessed by written response)
Reading (Grades Eleven and Twelve) 1.0 Word Analysis	
2.0 Reading Comprehension	17% (strand also assessed by written response)
3.0 Literary Response and Analysis	10% (strand also assessed by written response)
Total	100%

Written Composition

The GSE in Written Composition assesses writing standards for grades 9-10 and 11-12. Most of these standards are addressed by written-response items. Because the written composition multiple-choice items on the winter 2000 exam emphasized editing and revision, nearly all items used for the table below were judged to align most closely to the written and oral English language conventions standards.

Written Composition Content Strand	Percentage of multiple-choice items per content strand
Writing (Grades Nine and Ten)	
1.0 Writing Strategies	This strand is assessed by written response.
2.0 Writing Applications (Genres and Their Characteristics)	This strand is assessed by written response.
Written and Oral English Language Conventions	
1.0 Written and Oral English Conventions	96% (strand also assessed by written response)
Writing (Grades Eleven and Twelve)	
1.0 Writing Strategies	4% (strand also assessed by written response)
2.0 Writing Applications (Genres and Their Characteristics)	This strand is assessed by written response.
Total	100%

Mathematics

The Golden State Examinations in mathematics include first-year algebra and geometry, to be administered in spring 2000, and high school mathematics, administered in winter 2000.

First-year Algebra

The mathematics content standards do not separate the algebra standards into strands. Algebra 1 has been divided into the substrands shown below for convenience of presentation. Substrands 1, 2, and 3 each contain approximately one-third of the Algebra 1 strands.

First-year Algebra Content Area/Strand	Percentage of multiple-choice items per strand
Algebra 1	
Substrand 1 (standards 1-10)	53% (strand also assessed by written response)
Substrand 2 (standards 11-21)	14% (strand also assessed by written response)
Substrand 3 (standards 22-25.3)	6% (strand also assessed by written response)
Geometry*	12%
Probability and Statistics*	12%
Grade Seven: Algebra and Functions**	3%
Total	100%

*Items identified as addressing geometry and probability and statistics standards were included because they apply knowledge of algebra to those strands.
******A grade seven algebra and functions item that is foundational for Algebra I was included in the first-year algebra exam.

Geometry

The mathematics content standards do not divide the geometry standards into strands. Geometry has been divided into the substrands shown in the table below for convenience of presentation. Substrands 1, 2, and 3 each contain approximately one-third of the geometry strands.

Geometry Content Strand	Percentage of multiple-choice items per content strand
Geometry	
Substrand 1 (standards 1-7)	28% (strand also assessed by written response)
Substrand 2 (standards 8-14)	38% (strand also assessed by written response)
Substrand 3 (standards 15-22)	34% (strand also assessed by written response)
Total	100%

High School Mathematics

The GSE in High School Mathematics consists of a combination of four mathematics disciplines. The strands shown in the table below are the four disciplines that make up the examination.

High School Mathematics Content Area/Strand	Percentage of multiple-choice items per content area/strand
Algebra 1	30% (strand also assessed by written response)
Geometry	32% (strand also assessed by written response)
Algebra II	24% (strand also assessed by written response)
Probability and Statistics	14%
Total	100%

Science

Biology

In addition to standards that refer specifically to content, the biology standards include an investigation and experimentation strand that describes principles of investigation and experimentation. This strand is addressed by a number of items on the spring 2000 GSE in Biology, as indicated in the table below.

Biology Content Strand	Percentage of multiple-choice items per content strand
1. Cell Biology	24%
2–5. Genetics	26% (strand also assessed by lab task)
6. Ecology	22% (strand also assessed by lab task)
7–8. Evolution	8%
9–10. Physiology	6%
11. Investigation and Experimentation	14% (strand also assessed by lab task)
Total	100%

Chemistry

Like the biology standards, the chemistry standards include an investigation and experimentation strand. As the table below indicates, portions of the spring 2000 chemistry lab task and multiple-choice items address standards in this strand as well as content-specific standards.

Chemistry Content Strand	Percentage of multiple-choice items per content strand
1. Atomic and Molecular Structure	20%
2. Chemical Bonds	17% (strand also assessed by lab task)
3. Conservation of Matter and Stoichiometry	17% (strand also assessed by lab task)
4. Gases and Their Properties	5%
5. Acids and Bases	7% (strand also assessed by lab task)
6. Solutions	3% (strand also assessed by lab task)
7. Chemical Thermodynamics	11%
8. Reaction Rates	8%
9. Chemical Equilibrium	5% (strand also assessed by lab task)
10. Organic and Biochemistry	
11. Nuclear Processes	
12. Investigation and Experimentation	7% (strand also assessed by lab task)
Total	100%

Second-year Coordinated Science

Because the GSE in Second-year Coordinated Science includes items from biology, chemistry, physics, and earth science, these four disciplines serve as strands in the table below. Investigation and experimentation standards are also assessed in this exam.

Second-year Coordinated Science Content Areas	Percentage of multiple-choice items per content strand/area
Biology	33%
Chemistry	20% (strand also assessed by lab task)
Physics	16%
Earth Science	27% (strand also assessed by lab task)
Investigation and Experimentation	4% (strand also assessed by lab task)
Total	100%

Physics

Evaluation of alignment of the spring 2000 GSE in Physics included the six strands indicated in the table below.

Physics Content Strand	Percentage of multiple-choice items per content strand
1. Motions and Forces	28% (strand also assessed by lab task)
2. Conservation of Energy and Momentum	23% (strand also assessed by lab task)
3. Heat and Thermodynamics	9%
4. Waves	20% (strand also assessed by lab task)
5. Electronic and Magnetic Phenomena	20%
6. Investigation and Experimentation	This strand is assessed by lab task.
Total	100%

Spanish

Spanish Language

There are presently no standards for foreign language instruction in California. In the absence of state standards, the GSE in Spanish Language has been aligned to the *National Standards for Foreign Language Instruction*. The strands for those standards appear in the left column of the table below.

Spanish Language Strand	Percentage of multiple-choice items per content strand
1. **Communication**—Communicate in languages other than English	54% (strand also assessed by written response)
2. **Cultures**—Gain knowledge and understanding of other cultures	19%
3. **Connections**—Connect with other disciplines and acquire information	10%
4. **Comparisons**—Develop insight into the nature of language and culture	17% (strand also assessed by written response)
5. **Communities**—Participate in multilingual communities at home and around the world (not assessable on this test)	
Total	100%

Section Two: Validity

Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by the proposed uses of test results. The process of validation involves accumulating evidence to provide a sound scientific basis for the proposed interpretations of the scores. Evidence for validity can come from several sources. Two sources of validity evidence for the Golden State Examinations are discussed below.

Evidence Based on Content

Content evidence means that test items and the responses required by those items fully cover the intended content or curriculum. In California, the curriculum is defined by the statewide content standards. Content evidence may be verified by judges with content expertise who examine test items and confirm their alignment to standards. This method of alignment is noted in the *Standards for Educational and Psychological Testing*, Standard 1.7, which identifies expert opinion as one way of establishing content validity and describes guidelines for appropriate use of experts. Additional methods of establishing content validity can be used to verify the judgements of experts.

For the winter and spring 2000 Golden State Examinations, independent experts verified content validity by examining all test items and confirming their alignment to content standards. For the spring 2000 exams, these experts identified direct, explicit relationships between items and standards. To verify alignment, they required that the knowledge, information, activity, and skill described in each standard be addressed by the item. If each standard was directly addressed, and if the test as a whole demonstrated appropriate standards coverage, the test was judged to have content validity.

The content validity of the spring 2000 Golden State Examinations was established through direct alignment. This occurs when experts agree through consensus that items match the content of the standards. To establish alignment for the spring 2000 Golden State Examinations, a systematic process was used that included initial alignment of items to standards by GSE development teams followed by judgements of both internal reviewers and independent experts that items were directly aligned. Direct alignment was established by identifying a match between content of the standard and content of the item, including a match between the cognitive activity described in the standard and in the item; a match between the range of knowledge covered by the standard and by the item; and a balanced distribution of items across standards. When a test was found not to be fully aligned, as with the winter 2000 GSE in Written Composition, reviewers developed recommendations to ensure this examination’s alignment for its winter 2001 administration.

Evidence Related to Additional Types of Validity

As honors exams, the Golden State Examinations are designed to indicate high achievement among California's high school students. CDE is currently working with the California State University (CSU) system, which draws from the top 35 percent of

California's high school graduates, on studies exploring the use of the results from the GSE in Written Composition as predictors of student performance in CSU entry-level writing classes. CSU is considering the use of these results in lieu of those from the English Placement Test for placing students in these classes. CSU is also studying the use of results from the GSE in High School Mathematics as one criterion for placing students in entry-level mathematics classes. In addition, the University of California is studying the use of GSE results for students who achieve high honors on the written composition and high school mathematics exams as factors to be considered for UC admissions. Evidence from CSU and UC on the predictive accuracy of these examinations will provide CDE with data in addition to content evidence as it works to further ensure the validity of the Golden State Examinations.

Section Three—Reliability

Reliability is the general term used in educational measurement to describe the accuracy or precision of test scores. Accuracy is important because the more accurate test scores are, the more confidence there is in making inferences about the students' knowledge (e.g., in U.S. history) based on the scores. Accuracy of test scores can be improved by increasing the amount of information that tests provide about student performance.

Increasing Accuracy on Golden State Exams

The Golden State Examinations have been subjected to specific changes to increase the amount of information that these tests provide about student performance.

- The one 45-minute written-response section in history, economics, government/civics, reading/literature, and written composition has been changed to two 22-minute written responses. Two responses provide twice the information on students' content knowledge as one written response.
- Component rather than holistic scoring will be used for the GSE in Reading/Literature (winter 2000) and for the science lab tasks (spring 2000). With component scoring, tasks will be scored not as a whole but on several (e.g., four) dimensions or stages that make up each response. Four scores provide more information than one score and thereby increase accuracy.
- Students' multiple-choice and written-response scores have been combined onto a common scale combining these two scores into one. This allows the increase of accuracy from more information to be captured in one score known as the scale score.

Evidence Related to Additional Types of Reliability

Standard Error of Measurement. One way to look at accuracy is to look at the standard errors of measurement. The following table shows the conditional standard errors of measurement (SE) for scale scores at the cut points for the six performance levels. One way to use this table is to look at the SE for each exam at the cut point for performance level 4. If this value is low (i.e., lower than the other standard errors) then the exams are measuring most accurately at the point where the most important decisions are being made. That is, the exams are measuring most accurately where they are intended to measure. As can be seen on the following table, the exams are generally most accurate around the cut point for performance level 4. For example, algebra has a standard error of .23 at that cut point. The values for the standard errors at level 4 and level 5 (standard error = .22) are lower than the values at the other cut points. Therefore, algebra is measuring accurately at the point at which it needs to be most accurate.

Standard Errors for Performance Levels on the Golden State Examinations Spring 1999

Content	PL*=2	PL*=3	PL*=4	PL*=5	PL*=6
	SE**	SE	SE	SE	SE
Algebra	0.31	0.25	0.23	0.22	0.25
Geometry	0.26	0.23	0.23	0.26	0.27
U.S. History	0.29	0.21	0.23	0.25	0.25
Economics	0.30	0.23	0.21	0.22	0.22
Govt./Civics	0.25	0.22	0.22	0.24	0.24
Biology	0.37	0.31	0.31	0.31	0.31
Chemistry	0.32	0.29	0.29	0.31	0.32
Coord. Sci.	0.33	0.28	0.28	0.29	0.29
Physics	0.26	0.24	0.26	0.26	0.28
Spanish	0.26	0.23	0.23	0.25	0.28

* *This performance level cut point table identifies the standard errors at the cut point between each performance level and the performance level below it.*

** *Standard error*

The exams are generally most accurate around the cut point at performance level 4, and this is the performance level cut point separating students who receive recognition for their achievements from those who do not.

Figure 1 shows the conditional standard errors of measurement (SE) for algebra across scale scores (Ability) and performance levels (PL). As can be seen, algebra measures most accurately at the cut point for performance levels 4 and 5. Also, the exam is measuring most accurately the higher performing students (i.e., students performing above the mean ability level of 0).

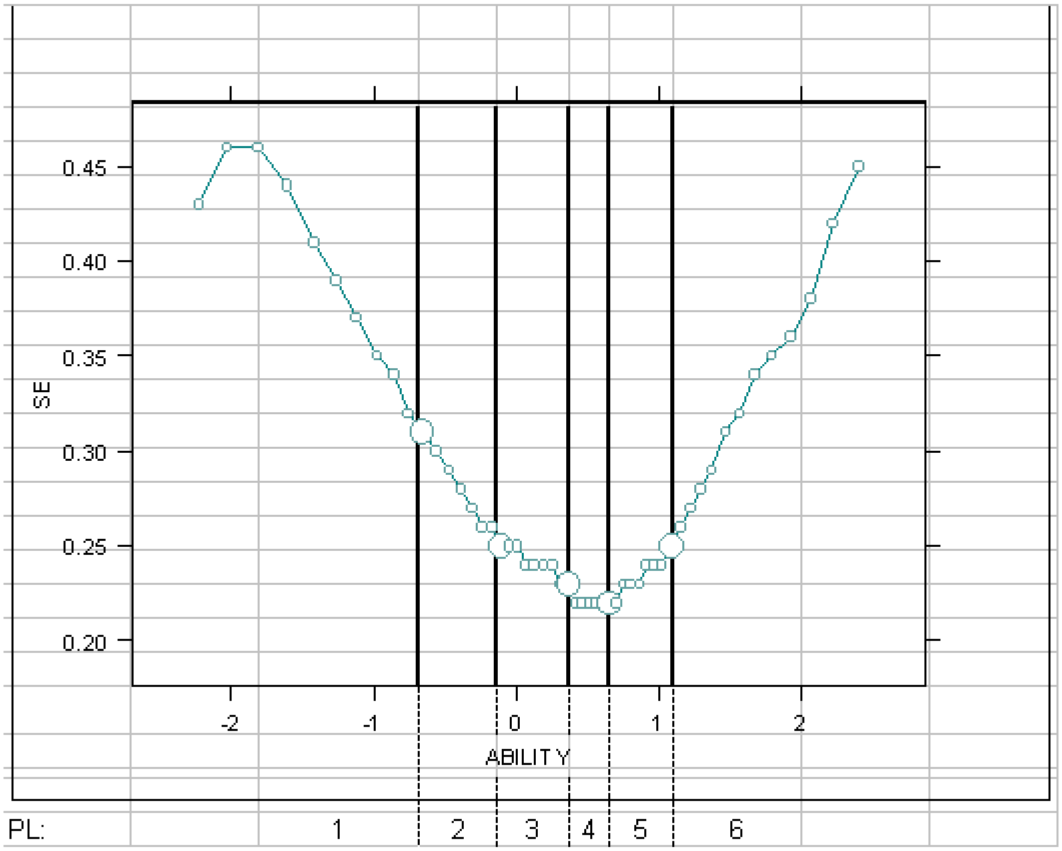


Figure 1. Conditional standard errors of measurement for algebra across scale scores and performance levels

Reliability Coefficients. The reliability coefficient is a kind of correlation. The values for reliability coefficients range from 0 to 1; 0 means no accuracy and 1 means perfect accuracy. Since all tests have error, reliability coefficients never reach 1. Reliability coefficients can be estimated from (1) the correlation of test scores from two parallel test forms, (2) the correlation of test scores from the same test form given on two different occasions (i.e., test/retest), and (3) the correlation of test scores from two halves of the same test (i.e., split halves). Spearman-Brown is one common method to calculate reliability coefficients using split halves.

The correlation of test scores from two halves of the same test is referred to as a measure of internal consistency. The advantage of using measures of internal consistency is that they can be estimated using only one administration of the test. Because of the convenience of a single test administration, measures of internal consistency are the most widely used procedures for estimating reliability coefficients. There are ways other than split halves to estimate reliability coefficients using internal consistency. One popular internal consistency procedure is Kuder-Richardson Formula 20 (KR20). KR20 is much more popular than using split halves. Further information about this and other aspects of reliability may be found in Appendix B.

Comparing GSE to Other Exams. Using KR20, Golden State Examinations were compared to two nationally recognized examinations which also measure high performing students, the Advanced Placement (AP) examinations and the Scholastic Aptitude Tests (SAT II).

When estimating reliability coefficients, the number of items on an examination affects the value. Tests with higher numbers of items will have higher reliability coefficients. All of the AP and SAT II examinations have at least 25 more items than the comparable Golden State Examinations. The Golden State Examinations are currently limited to 90 minutes of testing time. The number of multiple-choice items on the Golden State Examinations could be increase, but this may entail an increase in testing time.

The following chart shows the reliability coefficients for chemistry corresponding to the actual number of items on each test.

Subject	GSE		AP		SAT II	
	Number of Items	**KR-20****	**Number of Items**	**KR-20****	**Number of Items**	**KR-20****
Chemistry	35	0.82	75	0.93	84–85	.93–.94

In order to make a more accurate comparison of reliability coefficients (and thus the technical quality of the test items), the Spearman-Brown formula can be used to estimate reliability coefficients as if the tests were of the same length. The following charts show the reliability coefficients for the examinations calculated as if the AP and SAT II examinations had the same number of items as the Golden State Examinations.

Reliability Coefficients for Examinations in GSE, AP and SAT II (corrected for number of items)

Subject	GSE		AP		SAT II	
	Number of Items	**KR-20****	**Number of Items**	**KR-20****	**Number of Items**	**KR-20****
U.S. History	50	.89	50	.86	50	.85–.86
Biology	35	.76	.35	.82	35	.79–.88
Chemistry	35	.82	35	.84	35	.85–.87

Reliability Coefficients for Examinations in GSE and AP Only (corrected for number of items)

Subject	GSE		AP	
	Number of Items	KR-20**	Number of Items	KR-20**
Gov/Civics	50	.89	50	.87

Reliability Coefficients for Examinations in GSE and SAT II Only (corrected for number of items)

Subject	GSE		SAT II	
	Number of Items	KR-20**	Number of Items	KR-20**
Physics	35	.78	35	.83–.84
Written Comp	25	.70	25	.74–.79
Reading/Lit	35	.79	35	.78–.84
Spanish	40	.90	40	.83–.86

These comparisons support the conclusion that when corrected for the numbers of items, Golden State Examinations have comparable reliability coefficients to other national tests for high performing students.

Increasing Accuracy by Combining Multiple-choice and Written-Response Items. The following chart shows that when reliability coefficients are calculated for scale scores that combine multiple-choice (MC) and written-response (WR) items, the reliability coefficient for these scores improves over the reliability coefficients for multiple-choice items only.

Comparing GSE Reliability Coefficients with and without Written-response Scores

Subject	MC Only		MC + WR	
	No. of Items	Reliability	No. of Items	Reliability
Economics	50	.85-.86	52	0.86
US History	50	0.89	52	0.90
Govt./Civics	50	.86-.89	52	0.90
Biology	35	0.76	36	0.79
Chemistry	35	0.82	36	0.84
Coord. Sci.	35	0.76	36	0.84
Physics	35	0.78	36	0.81
Algebra	30	0.72	35	0.79
Geometry	30	0.77	35	0.82
Spanish	40	0.90	42	0.91

Section Four – Conclusion

The California Department of Education (CDE) is in an ongoing process to ensure that the Golden State Examinations are aligned to the statewide content standards, and are valid and reliable. The reviews discussed in this report have identified areas in which GSE alignment, validity, and reliability are soundly established and areas in which they need to be strengthened. The standards-alignment reviews have established that the spring 2000 Golden State Examinations are aligned to standards. For the reading/literature and written composition exams, which are next administered in winter 2001, GSE will use review panel recommendations to align the tests for the winter 2001 administration.

The alignment reviews have confirmed the validity of the spring 2000 Golden State Examinations. The CDE is using the results of these reviews to ensure the validity of the winter 2001 reading/literature and written composition tests. GSE is participating in studies with the California State University and the University of California to investigate means of further strengthening the validity of the Golden State Examinations.

The technical reviews have established that the Golden State Examinations provide accurate scores at the level critical for identifying students who qualify for honors recognition. To improve reliabilities, the CDE has implemented a number of measures. These include increasing the number of written responses required of students on certain exams from one to two, replacing holistic with component scoring on written-response items and lab tasks, and converting students' multiple-choice and written-response scores to a common scale.

The technical analyses suggest that additional measures could be undertaken to increase GSE reliabilities. These include increasing the number of multiple-choice items on the examinations, which may entail an increase in testing time. These measures would require modifications in GSE test designs but would lead to greater accuracy in the test scores. This may be an opportune time to implement such revisions in the Golden State Examinations.

Appendix A

Alignment of Spring 2000 GSE Examinations to Content Standards
Participants in Independent Reviews

September – October 1999 Meetings

History/Social Science

Lucy Barber, University of California, Davis
Carolita Carr, San Juan Unified School District
Stanley Clark, California State University, Bakersfield
Kurt Dearie, Carlsbad Unified School District
Paul Garcia, Fresno Unified School District
Tom Jacoubowsky, Sequoia Union High School District
Deme Larson, Keppel Union Elementary School District
Elizabeth Mitchell, California School Boards Association
Bill Palmer, Lodi Unified School District
David Pava, New Haven Unified School District
Linda Tubach, Los Angeles Unified School District
Gary Wexler, William S. Hart High School District

Language Arts

Dodie Andersen, Chino Unified School District
Kathleen Coughlin, Sequoia Union School District
Lynne Culp, Los Angeles Unified School District
Maria Gautreau, West Covina Unified School District
Cathryn Geyer, Lodi Unified School District
Linda Menville, Imperial County Office Education
Elizabeth Mitchell, California School Board Association
Aaron Spain, Morgan Hill Unified School District

Math

Ruth Asmundson, Davis Joint Unified School District
John Briggs, Central Union High School District
Liz Brookins, University of California, San Diego
Steve Cook, Rim of the World Unified School District
Priscilla Cox, California School Boards Association
Mercidita Del Rosario, Kern Union High School District
Dorothy Haas, Los Alamitos Unified School District
Grace Hutchings, Los Angeles Unified School District
Louise Iverson, Gold Train Union High School District

Kathy Moffat, California State PTA
Anuar Shalash, Alhambra City High School District
Sue Stickel, Elk Grove Unified School District
Carol Treglio, San Diego Unified School District

Science

Ruth Asmundson, Davis Unified School District
Helen Finks, New Haven Unified School District
Barbara Howe, California State PTA
Don Hubbard, Retired
Randy Malandro, Lodi Unified School District
Kathy Moffat, California State PTA
Jim Postma, Chico Unified School District
Wendell Potter, University of California, Davis
Michael Rios, Montebello Unified School District
Todd Samet, Tamalpias Union High School District
Laurie Schonert, Rim of the World Unified School District
Jerry Valadez, Fresno Unified School District
Don Yost, Folsom/Cordova Unified School District
Kendall Zoller, San Juan Unified School District

March 23, 2000 Meetings

History/Social Science

Jeff Dellis, Nevada Joint Union High School District
Krista Dornbush, Natomas Unified School District
Susan Harmon, Huntington Beach Unified School District
Janet Landfried, Redlands Unified School District
Scott Luhn, Stockton Unified School District

Language Arts

Laurie Brooke, Lodi Unified School District
Karen Hayashi, Elk Grove Unified School District
Bruce Holden, Nevada Joint Union High School District
Micki Sanders, Sacramento Unified School District
Laura Watson, Elk Grove Unified School District

Math

Ruth Asmundson, Davis Joint Unified School District
Marin Beechen, Chaffey Unified School District
Karen Cliffe, Sweetwater Union High School District
Toni Smith, Rim of the World Unified School District

Science

Ann Akey, Sequoia Union High School District
Kathy Iverson, Huntington Beach Unified School District
Joe Monaco, Redlands Unified School District
Carol Anne Piehl, Sequoia Union High School District
Laurie Schonert, Rim of the World Unified School District
Ellen Vasta, Elk Grove Unified School District

Appendix B

Reliability

Reliability is the general term used in educational measurement to describe the accuracy or precision of test scores. Reliability is important because the more accurate a test score the more confidence there is in making inferences about an examinee's content knowledge based on the test score. Accuracy of test scores can be improved by increasing the amount of information that tests elicit from examinees. This paper examines different ways accuracy can be described and estimated. Some of these include; (1) standard errors (or error variance) of measurement, (2) reliability coefficients, (3) misclassification errors, and (4) for written-response items (e.g., write an essay on the causes of the Civil War) indexes of rater consistency.

Standard Error of Measurement

The standard error of measurement represents the standard deviation of a hypothetical set of repeated measures on a single examinee. That is, if an examinee could take the same exam over and over (forgetting the experience before each new administration) the standard error represents the variability of these hypothetical test scores in standard units. The mean of this hypothetical distribution of scores is referred to in classical test theory as an examinee's true score. Error is the difference between an examinee's actual score and an examinee's true score.

The standard error of measurement is very useful because it shows how much error exists around actual test scores. The standard error can be used to create score intervals around test scores (e.g., the test score is in an interval that is one standard error above the score and one standard error below the score). Although this interval provides a notion about the degree of error, it is difficult to understand and explain.

The standard error of measurement (using the score interval as a measure of accuracy) can be interpreted to mean that other comparable intervals will capture an examinee's true score a certain percent of the time. That is, a score interval of one standard error above an examinee's score and one standard error below the score captures the examinee's true score 68% of the time. If two standard errors are used, other comparable intervals (i.e., two standard errors above and two standard errors below) will capture an examinee's true score 95% of the time. How close an examinee's actual test score is to the true score is unknown. An examinee's actual test score may be close or far away from the true score.

Typically one estimate (i.e., one value for the standard error) is derived for the whole range of test scores that are possible for a test (e.g., all the possible number correct scores for the second grade mathematics test). However, a standard error can be computed for each score that exists for a particular test (e.g., each number correct score for a second grade mathematics test). These estimates are called the conditional standard errors of measurement and show the relative accuracy of each test score. That is, these estimates show that some scores measure more accurately than other scores. Scores on norm-referenced standardized tests (e.g., the Stanford 9 Achievement Test) generally are more accurate for middle performing students than for high performing and low performing students. Figures 1 through 10 in this appendix show the conditional standard errors of measurement for several of the Golden State Examinations.

It is often useful to know which scores measure most (and least) accurately. For example, if scores above or below a certain cut point (i.e., a particular test score) are used to make important decisions about students, it is important to know about the accuracy of the cut point (i.e., test score).

The Reliability Coefficient

The reliability coefficient is a kind of correlation. Thus, the values for reliability coefficients range from 0 to 1; 0 means no accuracy and 1 means perfect accuracy. Since all tests have error, reliability coefficients never reach 1. Reliability coefficients can be estimated from (1) the correlation of test scores from two parallel test forms, (2) the correlation of test scores from the same test form given on two different occasions (i.e., test/retest), and (3) the correlation of test scores from two halves of the same test (i.e., split halves). Spearman-Brown is one common method to calculate reliability coefficients using split halves. It should be noted that these three methods of estimating reliability coefficients are actually three different aspects of precision.

The correlation of test scores from two halves of the same test is referred to as a measure of internal consistency. The advantage of using measures of internal consistency is that they can be estimated using only one administration of the test. Because of the convenience of a single test administration, measures of internal consistency are the most widely used procedures for estimating reliability coefficients. There are ways other than split halves to estimate reliability coefficients using internal consistency. One popular internal consistency procedure is Kuder-Richardson Formula 20 (KR20). KR20 is much more popular than using split halves.

One weakness of the reliability coefficient is that the same test can produce different reliability coefficients when administered to different examinees under different conditions. That is, the value of the reliability coefficient is sample-dependent. It is especially dependent on the heterogeneity of the examinees (i.e., the degree of difference between examinees on the content being measured).

Tables 1 and 2 in this appendix show the reliability coefficients for Golden State Examinations (GSE), Advanced Placement (AP) examinations, and the Scholastic Aptitude Tests (SAT II). Table 2 shows that when reliability coefficients are calculated for scale scores that combine multiple-choice and written-response items the reliability coefficient for these scores improves over the reliability coefficients for multiple-choice items only. SAT II scores are not included because SAT II does not include written-response items on the examinations, except for written composition.

Increasing the Value of the Reliability Coefficient Using KR20

Accuracy, as stated, can be improved by increasing the amount of information that tests elicit from examinees. When using internal consistency procedures (i.e., KR20) the value of the reliability coefficient can be increased by increasing the number of items. KR20 for multiple-choice tests can also be increased two other ways. (1) Increase the test score variance across examinees. Test score variance means that the scores for examinees of different abilities are different. The more that scores vary (i.e., the scores are different from each other) the greater the variance. (2) Make items more homogenous and reduce the item variance. The logic is that for a test score to be accurate examinees who know most about a content area (e.g., language arts) should be consistent in their correct responses, and examinees that know the least should be consistent in their incorrect responses. The response pattern for examinees of different achievement levels should be consistent for their level and not vary greatly from item to item.

As an estimate of internal consistency, KR20 looks at the ratio between score variance and item variance. As score variance goes up and item variance goes down the reliability coefficient increases in value. Since KR20 is dependent on homogeneity of items it will underestimate reliability if items measure very different aspects of a content area.

Increase the Number of Items. The more items a test contains the more opportunity an examinee has to demonstrate accurately what the examinee knows about a particular content area (e.g., language arts). If a test has only one multiple-choice item, for example, an examinee who knows little about the content may get the item right because (1) it happens to be a question about the one thing with which the examinee is familiar, or (2) the examinee guesses correctly. An examinee who is very knowledgeable about the content may get the item wrong because (1) it happens to be a question about the one thing with which the examinee is not familiar, or (2) the examinee accidentally marks the wrong option. ncreasing the number of items increases the opportunities for the examinee to demonstrate what the examinee really knows. That is, the score is a more accurate reflection of what the examinee knows.

Increase Test Score Variance. One way to increase the test score variance (and increase the reliability coefficient) is to use items with high point biserial correlations. The point biserial correlation is the correlation between a dichotomous variable and a continuous variable. In this case it is the correlation between the item which is dichotomous (i.e., 0=wrong and 1=correct) and the total test score. The point biserial correlation is often used to determine how well an item discriminates between those who know most about the content area (e.g., language arts) and those who know the least.

Reduce Item Variance. One way to reduce the item variance (and increase the reliability coefficient) is to make all the items as homogenous as possible. The logic is that examinees who do well on a particular type of item (e.g., a multiple-choice spelling item) continue to do well across lots of items of the same type and vice versa. There is internal consistency of examinee responses.

One way to increase item variance (and depress the reliability coefficient) is to use items that measure different aspects of a content area. For example, language arts is measured using spelling items, reading items, and grammar items.[1]

Since the number of items on examinations affects the value of the reliability coefficient, tests with higher numbers of items will have higher reliability coefficients. All of the AP and SAT II examinations have at least 25 more items than the comparable Golden State Examinations. Thus, in order to make a more accurate comparison the Spearman-Brown formula can be used to estimate reliability coefficients as if the tests were of the same length. Tables 3 and 4 in this appendix show the reliability coefficients for the examinations calculated as if the AP and SAT II examinations had the same number of items as the Golden State Examinations.

Misclassification Error

Misclassification error is an estimate that looks at how accurately examinees are classified. For example, an examinee's true percentile score for a particular test (e.g., second-grade mathematics) is 25 (i.e., the 25th percentile) and the reliability coefficient for this test is .9. In this case, how often will an examinee be identified within 5 (or 10) percentile points of the true score (i.e., the 25th percentile)? That is, how often (i.e., what proportion of the time) will an examinee with a true score at the 25th percentile have an actual test score between the 20th and 30th percentiles (or between the 15th and 35th percentiles)? For a test with .9 reliability, a student will have scores between the 20th and 30th percentiles 40 percent of the time. Therefore, 60 percent of the time a student will have scores outside this range or be misclassified 60 percent of the time. For a test with .9 reliability, a student will have scores between the 15th and 35th percentiles 70 percent of the time and be misclassified 30 percent of the time.

Rater Consistency on Written-response Items

Indexes of rater consistency estimate how accurately different raters score the same student's response. One estimate is the correlation of the scores from different raters for common student responses. It is sometimes assumed that written-response items are less reliable than multiple-choice items because there is rater error and because there are fewer items (because of time limitations). However, a test composed solely of written-response items would be much more reliable than a multiple-choice test that contained an equal number of items. One can generally learn much more about an examinee's knowledge of a particular content area (e.g., U.S. history) from one written-response item than from one multiple-choice item. And, one can learn much more about an examinee's knowledge of a particular content area (e.g., U.S. history) from fifty written-response items than from fifty multiple-choice items.

[1] Using items that measure different aspects of a content area (i.e., using spelling items, reading items, and grammar items to measure language arts) will improve validity but may decrease reliability.

It is not true that multiple-choice items are free from error. Multiple-choice items have error, including such things as guessing (i.e., guessing the right answer when the examinee really did not know the correct answer) and marking the wrong option (i.e., the examinee knew the correct answer but accidentally marked the wrong option). There are correction-for-guessing formulas that try to account for guessing by adjusting scores and item statistics that calculate a guessing parameter. However, there are more types of error than guessing and they are not all accounted for except in the general sense of error (i.e., the standard error of measurement).

For written-response items, rater error can be calculated and its effect on overall accuracy can be estimated. In this way it can be determined whether the best way to reduce error is to increase the number of items or improve rater accuracy and the costs (e.g., in terms of money and time) associated with each. However even with rater error (that can be estimated), the reliability (i.e., accuracy) will usually be greater for written-response items than for an equal number of multiple-choice items.

ALGEBRA SPRING 1999

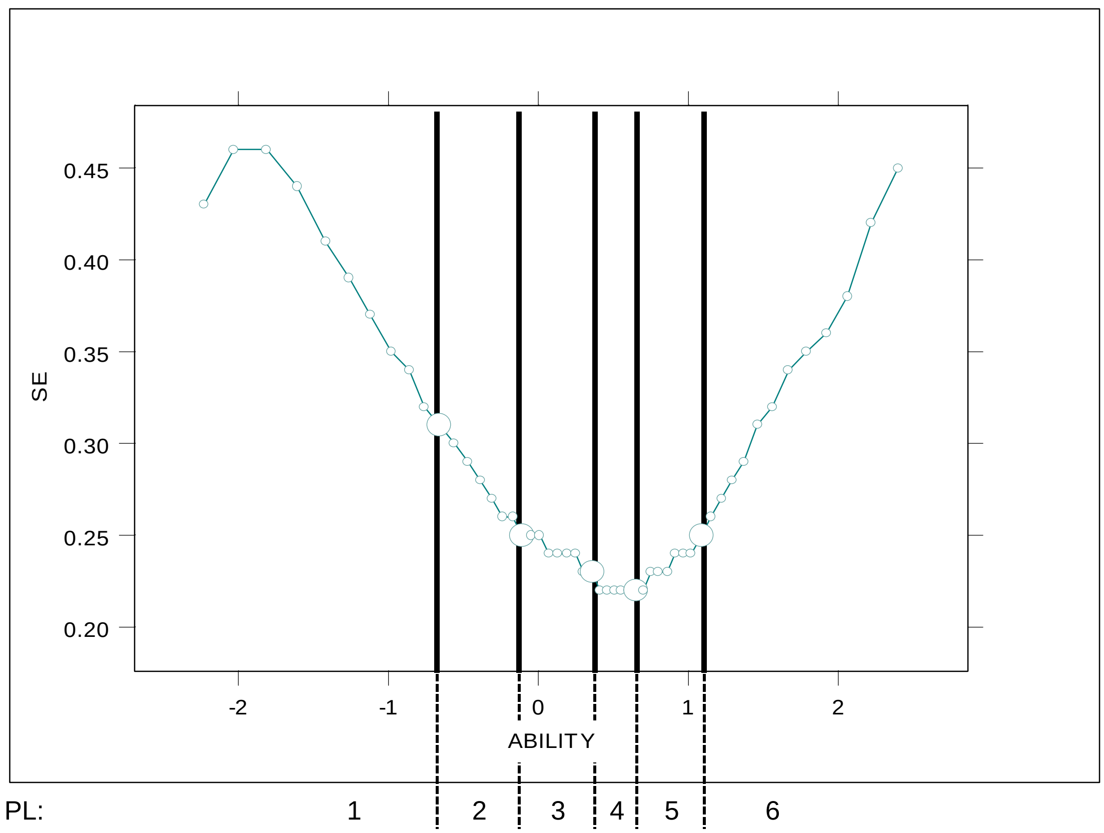


Figure 1. Conditional standard errors of measurement for algebra across scale scores and performance levels

GEOMETRY SPRING 1999

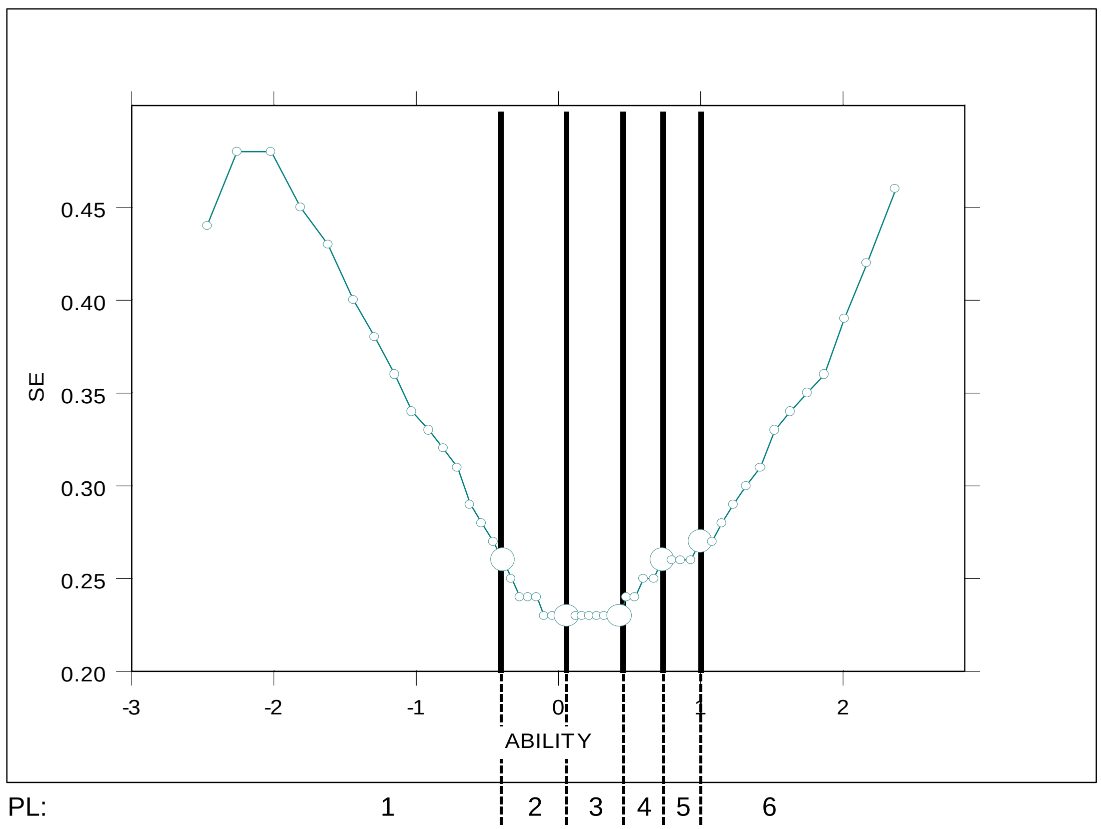


Figure 2. Conditional standard errors of measurement for geometry across scale scores and performance levels

U.S. HISTORY SPRING 1999

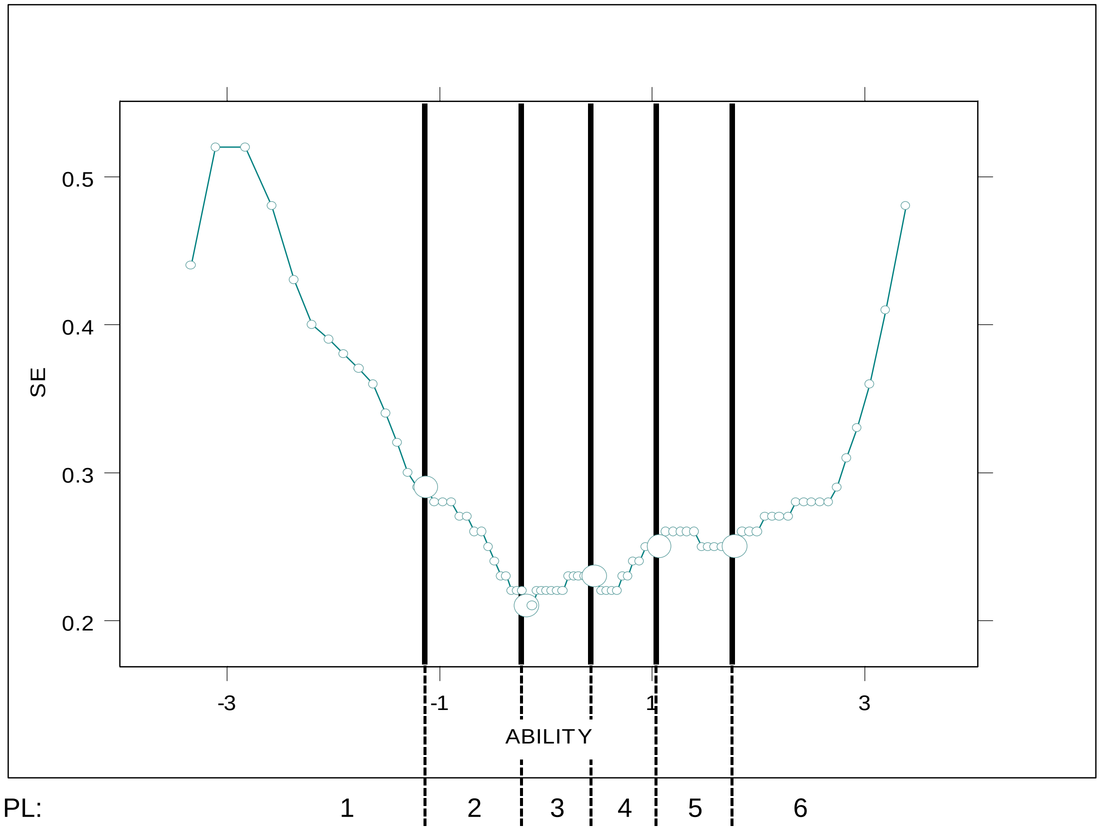
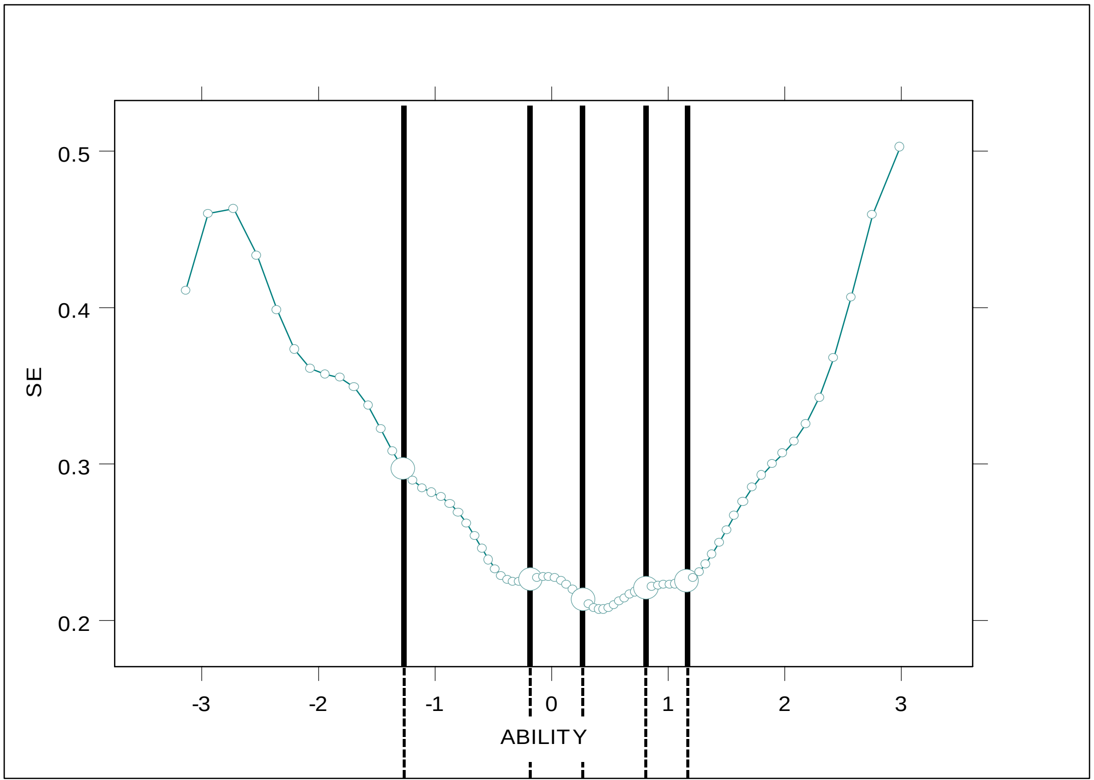


Figure 3. Conditional standard errors of measurement for history across scale scores and performance levels

ECONOMICS SPRING 1999



PL: 1 2 3 4 5 6

Figure 4. Conditional standard errors of measurement for economics across scale scores and performance levels

GOVERNMENT/CIVICS SPRING 1999

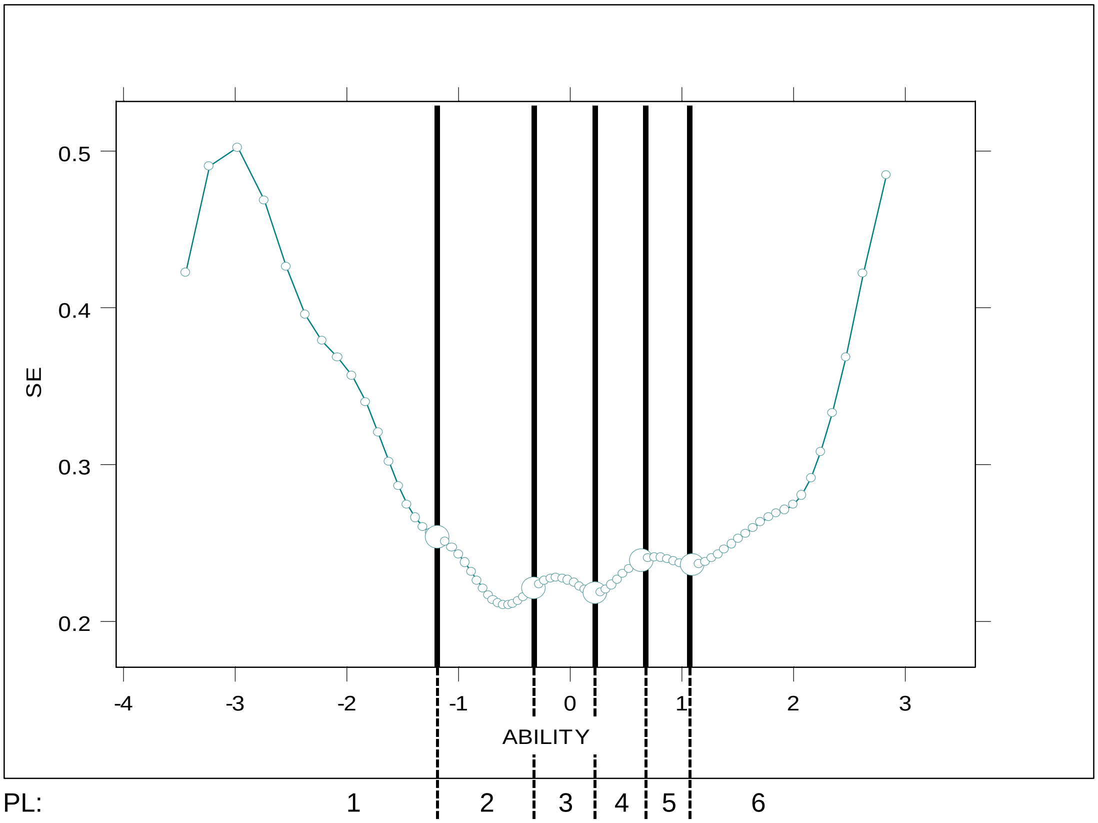


Figure 5. Conditional standard errors of measurement for govt./civics across scale scores and performance levels

BIOLOGY SPRING 1999

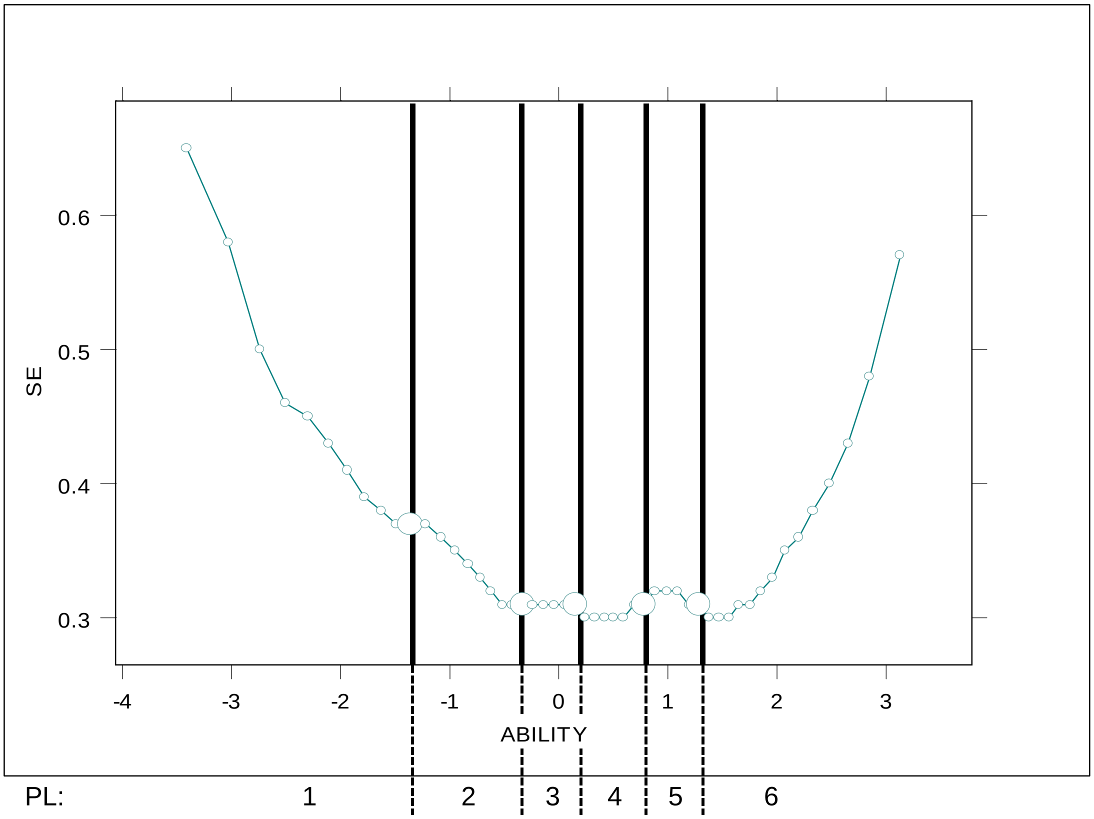


Figure 6. Conditional standard errors of measurement for biology across scale scores and performance levels

CHEMISTRY SPRING 1999

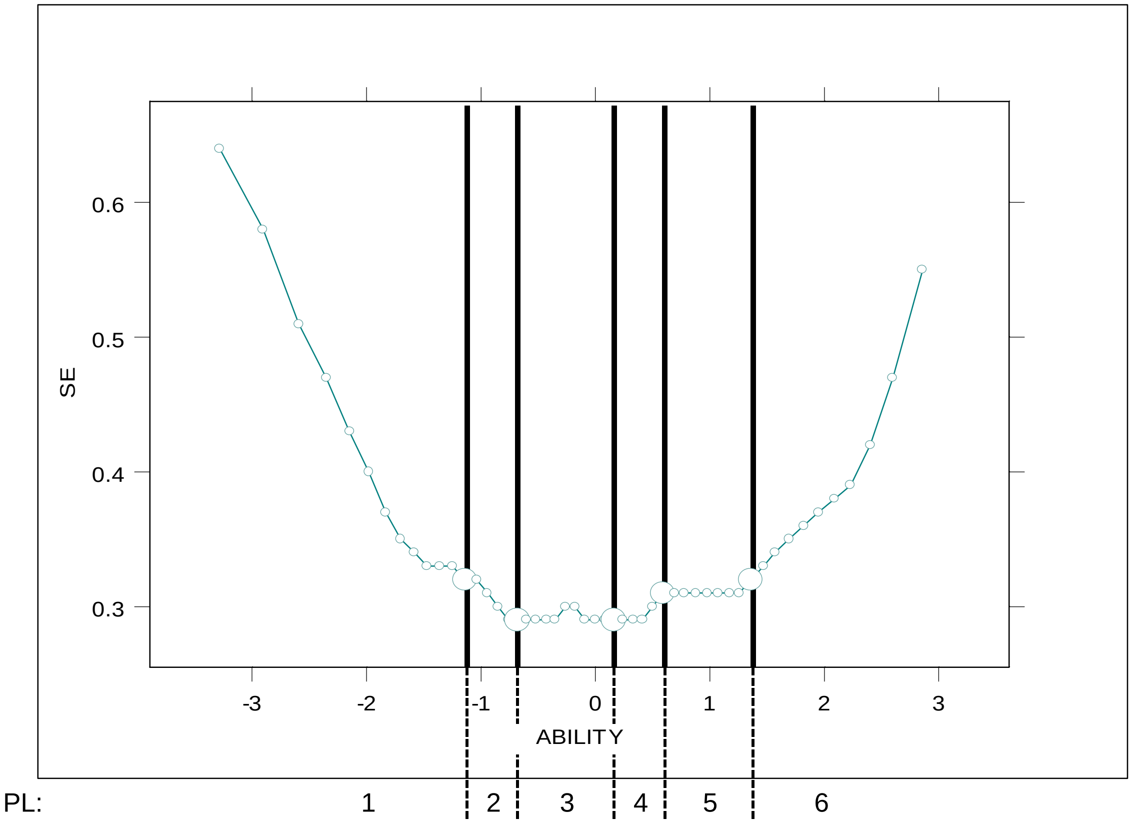


Figure 7. Conditional standard errors of measurement for chemistry across scale scores and performance levels

COORDINATED SCIENCE SPRING 1999

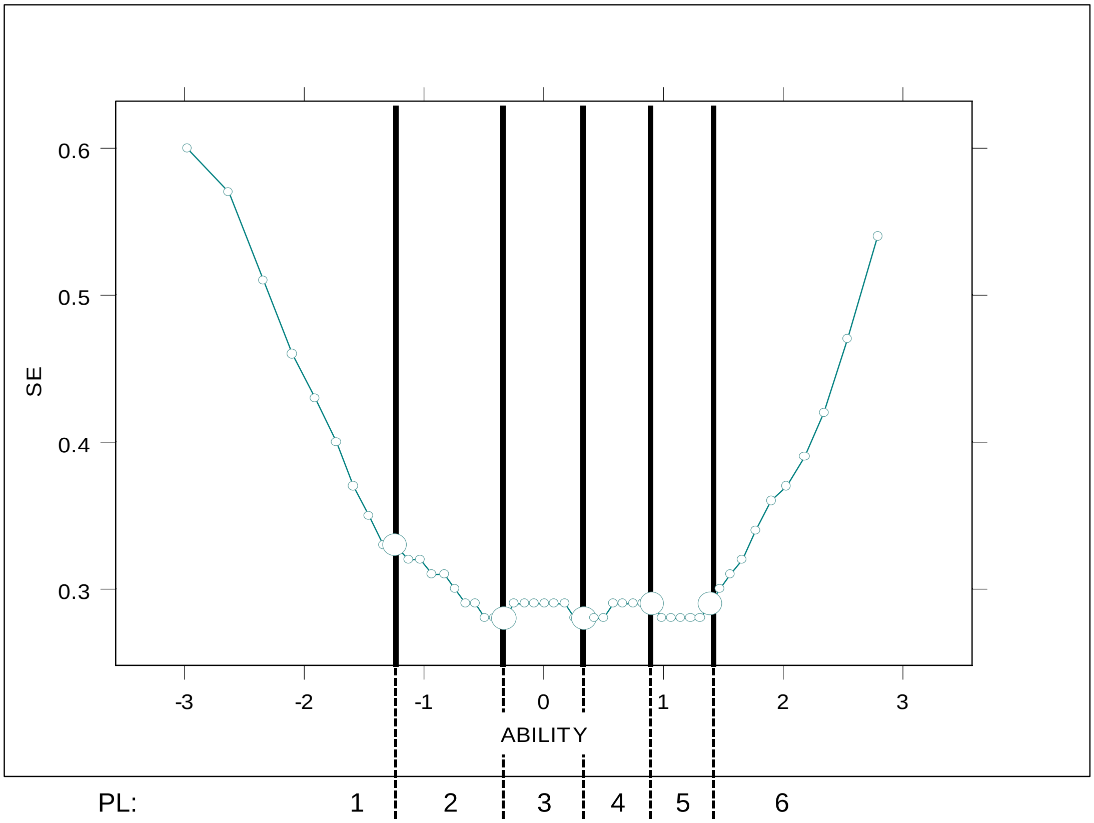


Figure 8. Conditional standard errors of measurement for coordinated science across scale scores and performance levels

PHYSICS SPRING 1999

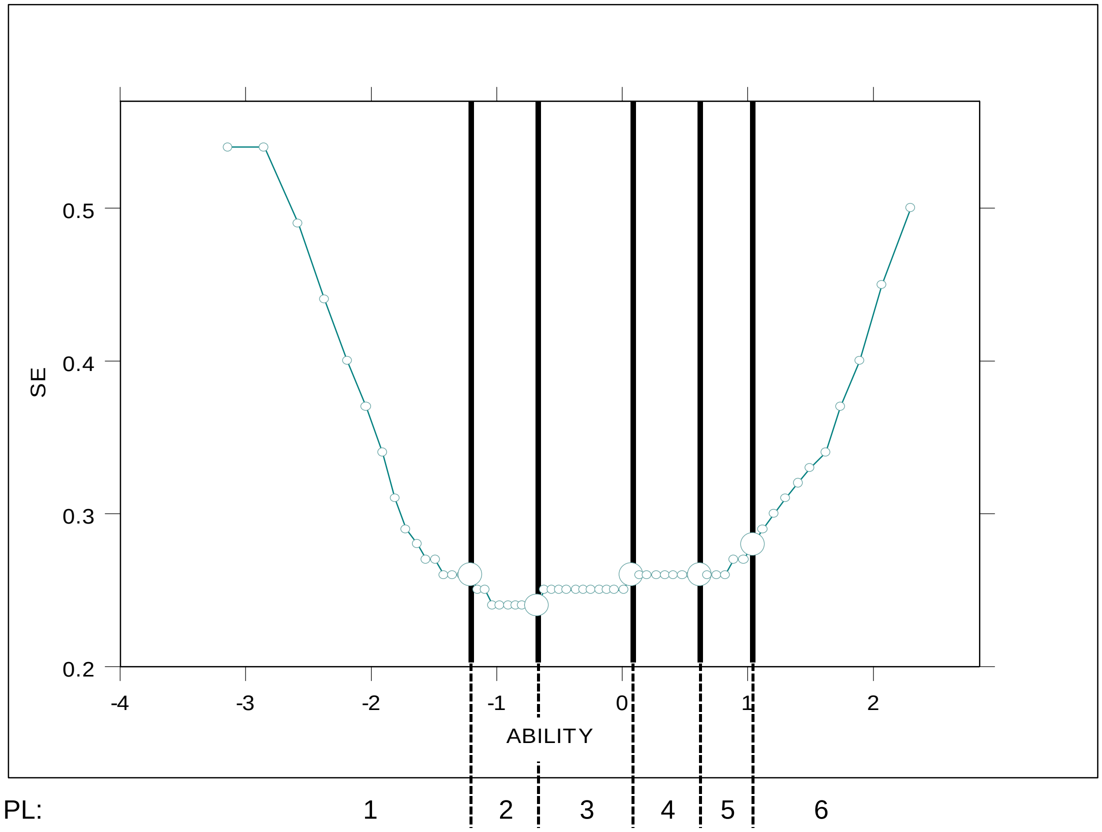


Figure 9. Conditional standard errors of measurement for physics across scale scores and performance levels

SPANISH (YEAR 2 ONLY) SPRING 1999

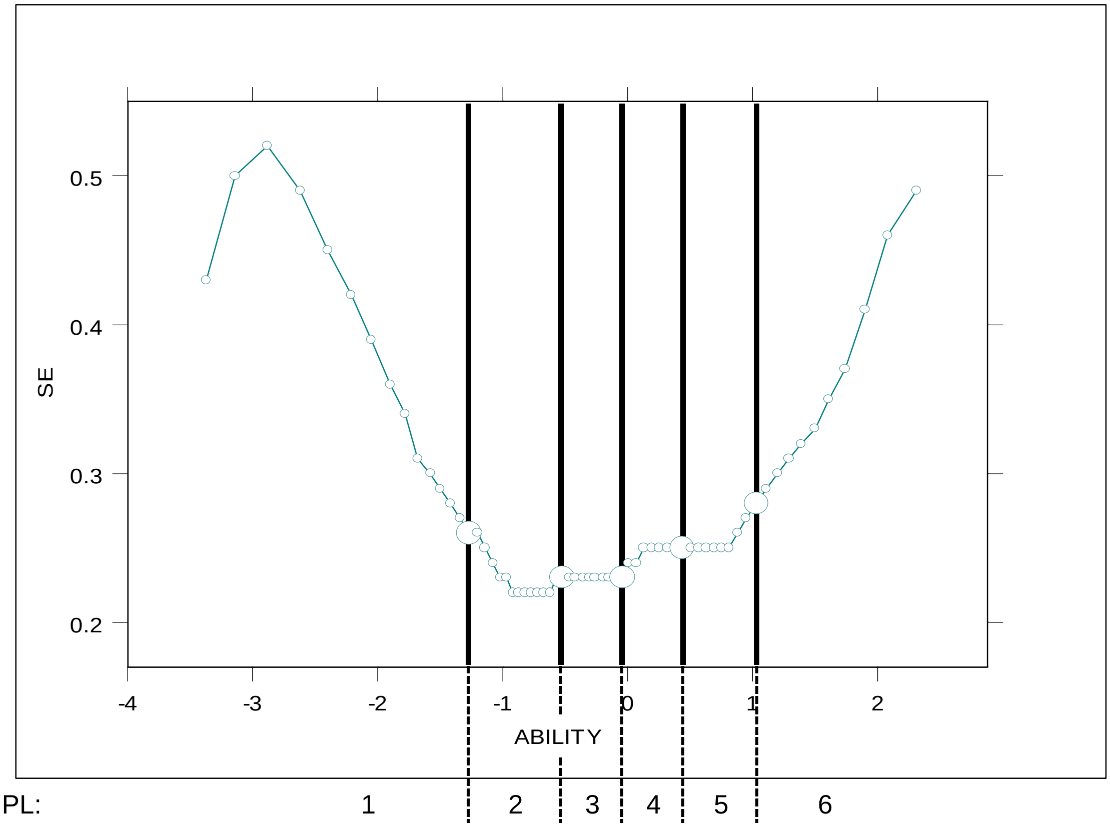


Figure 10. Conditional standard errors of measurement for Spanish across scale scores and performance levels

Table 1

Reliability Coefficients for Golden State Examinations, Advanced Placement, and SAT II

Multiple Choice

Subject	GSE		AP		SAT II	
	Number of Items	**KR-20***	**Number of Items**	**KR-20***	**Number of Items**	**KR-20***
Economics	50	.85–.86				
Economics Micro			60	0.90		
Economics Macro			60	0.89		
Government/Civics	50	.86–.89	60	0.88		
US History	50	0.89	80	0.89	88–90	.91–.92
Reading/Literature	35	0.79			60–62	.86–.90
Written Composition	25	0.70			59–60	.87–.90
English Language			55	0.86		
English Literature			55	0.86		
Algebra	30	0.72				
Geometry	30	0.77				
High School Math	30	0.78				
Calculus AB			45	0.89		
Calculus BC			45	0.88		
Math 1C					50	.86–.88
Math IIC					50	.88–.92
Biology	35	0.76	120	0.94	95	.91–.95
Chemistry	35	0.82	75	0.93	84–85	.93–.94
Coordinated Science	35	0.76				
Physics	35	0.78			75	.91–.92
Physics B			70	0.90		
Physics C Mechanics			35	0.86		
Physics Electricity			35	0.81		
Spanish	40	0.90			84–85	.91–.93
Spanish Language			90	0.88		
Spanish Literature			65	0.84		

• *Kuder-Richardson Formula 20 for computing a reliability coefficient.*

Table 2

Reliability Coefficients for Golden State Examinations and Advanced Placement

Multiple Choice + Written Response

Subject	GSE		AP	
	Number of Items	**Reliability**	**Number of Items**	**Reliability**
Economics	52	0.86		
Economics Micro			63	.91–.94
Economics Macro			63	.90–.94
Government/Civics	52	0.90	64	.87–.93
US History	52	0.90	83	.88–.92
Algebra	32	0.79		
Geometry	32	0.82		
Calculus AB			51	.93–.97
Calculus BC			51	.92–.96
Biology	36	0.79	124	.93–.96
Chemistry	36	0.84	81	.94–.98
Coordinated Science	36	0.84		
Physics	36	0.81		
Physics B			77	.93–.97
Physics C Mechanics			38	.90–.94
Physics C Electricity			38	.88–.94
Spanish	42	0.91		
Spanish Language			94	0.92
Spanish Literature			68	0.85

- *Kuder-Richardson Formula 20 for computing a reliability coefficient.*

Table 3

Reliability Coefficients for Golden State Examinations, Advanced Placement, and SAT II

The Advanced Placement and SAT II exams are adjusted to show the reliability coefficients if Advanced Placement and SAT II had the same number of items as GSE

Multiple Choice

Subject	GSE		AP		SAT II	
	Number of Items	**KR-20***	**Number of Items**	**KR-20***	**Number of Items**	**KR-20***
Economics	50	0.85				
Economics Micro			50	0.89		
Economics Macro			50	0.87		
Government/Civics	50	0.89	50	0.87		
US History	50	0.89	50	0.86	50	.85–.86
Reading/Literature	35	0.79			35	.78–.84
Written Composition	25	0.70			25	.74–.79
English Language			25	0.74		
English Literature	35	0.79	35	0.80	35	.78–.84
Algebra	30	0.72				
Geometry	30	0.77				
High School Math	30	0.78				
Calculus AB			30	0.84		
Calculus BC			30	0.83		
Math 1C					30	.79–.81
Math IIC					30	.81–.87
Biology	35	0.76	35	0.82	35	.79–.88
Chemistry	35	0.82	35	0.84	35	.85–.87
Coordinated Science	35	0.76				
Physics	35	0.78			35	.83–.84
Physics B			35	0.82		
Physics C Mechanics			35	0.86		
Physics Electricity			35	0.81		
Spanish	40	0.90			40	.83–.86
Spanish Language			40	0.77		
Spanish Literature			40	0.76		

• *Kuder-Richardson Formula 20 for computing a reliability coefficient.*

Table 4

Reliability Coefficients for Golden State Examinations and Advanced Placement

The Advanced Placement and SAT II exams are adjusted to show the reliability coefficients if Advanced Placement and SAT II had the same number of items as GSE

Multiple Choice + Written Response

Subject	GSE		AP	
	Number of Items	**Reliability**	**Number of Items**	**Reliability**
Economics	52	0.86		
Economics Micro			52	0.89
Economics Macro			52	0.87
Government/Civics	52	0.90	52	0.87
US History	52	0.90	52	0.86
Algebra	32	0.79		
Geometry	32	0.82		
Calculus AB			32	0.85
Calculus BC			32	0.83
Biology	36	0.79	36	0.82
Chemistry	36	0.84	36	0.84
Coordinated Science	36	0.84		
Physics	36	0.81		
Physics B			36	0.81
Physics C Mechanics			36	0.85
Physics C Electricity			36	0.80
Spanish	42	0.91		
Spanish Language			42	0.77
Spanish Literature			42	0.76

- *Kuder-Richardson Formula 20 for computing a reliability coefficient.*

Appendix C

Ethnic Distribution of Participants in 1999 Golden State Examinations

(Totals reflect most recent total of students)

WINTER 1999 GOLDEN STATE EXAMINATIONS

# of Students per Primary Ethnic Group

Content Area	(blank/ multiple)	African-American	Native American	Asian-American	Filipino-American	Hispanic/ Latino	Pacific Islander	White	Totals
Economics	3478	1943	293	4771	1320	7837	326	16126	36094
	9.6%	*5.4%*	*0.8%*	*13.2%*	*3.7%*	*21.7%*	*0.9%*	*44.7%*	*100%*
Government/Civics	3859	2078	343	4932	1339	8476	363	16434	37824
	10.2%	*5.5%*	*0.9%*	*13.0%*	*3.5%*	*22.4%*	*1.0%*	*43.4%*	*100%*
H.S. Math	3698	1186	207	8881	1694	5186	241	14729	35822
	10.3%	*3.3%*	*0.6%*	*24.8%*	*4.7%*	*14.5%*	*0.7%*	*41.1%*	*100%*
Reading & Lit.	6847	3763	555	9054	2550	15501	634	30416	69320
	9.9%	*5.4%*	*0.8%*	*13.1%*	*3.7%*	*22.4%*	*0.9%*	*43.9%*	*100%*
Written Comp.	10088	4833	769	12177	3539	20319	830	41424	93979
	10.7%	*5.1%*	*0.8%*	*13.0%*	*3.8%*	*21.6%*	*0.9%*	*44.1%*	*100%*

SPRING 1999 GOLDEN STATE EXAMINATIONS

# of Students per Primary Ethnic Group

Content Area	(blank/ multiple)	African-American	Native American	Asian-American	Filipino-American	Hispanic/ Latino	Pacific Islander	White	Totals
Algebra	17657	9921	1248	24313	7271	41813	1649	77036	180908
	9.8%	*5.5%*	*0.7%*	*13.4%*	*4.0%*	*23.1%*	*0.9%*	*42.6%*	*100%*
Biology	10492	6554	881	17197	4909	25931	1154	51604	118722
	8.8%	*5.5%*	*0.7%*	*14.5%*	*4.1%*	*21.8%*	*1.0%*	*43.5%*	*100%*
Chemistry	7493	3434	530	15237	3428	13554	633	34386	78695
	9.5%	*4.4%*	*0.7%*	*19.4%*	*4.4%*	*17.2%*	*0.8%*	*43.7%*	*100%*
Coordinated Science	2166	2252	174	2824	1162	7411	304	9762	26055
	8.3%	*8.6%*	*0.7%*	*10.8%*	*4.5%*	*28.4%*	*1.2%*	*37.5%*	*100%*
Economics	3637	2266	332	5200	1467	9387	392	16488	39169
	9.3%	*5.8%*	*0.8%*	*13.3%*	*3.7%*	*24.0%*	*1.0%*	*42.1%*	*100%*
Geometry	11311	5803	768	19018	5128	23344	1052	54351	120775
	9.4%	*4.8%*	*0.6%*	*15.7%*	*4.2%*	*19.3%*	*0.9%*	*45.0%*	*100%*
Government/Civics	3119	1849	289	4250	1317	7730	350	15020	33924
	9.2%	*5.5%*	*0.9%*	*12.5%*	*3.9%*	*22.8%*	*1.0%*	*44.3%*	*100%*
Physics	2952	1175	224	7998	1526	4449	279	13877	32480
	9.1%	*3.6%*	*0.7%*	*24.6%*	*4.7%*	*13.7%*	*0.9%*	*42.7%*	*100%*
Spanish	6543	3275	488	11086	3015	27839	669	36875	89790
	7.3%	*3.6%*	*0.5%*	*12.3%*	*3.4%*	*31.0%*	*0.7%*	*41.1%*	*100%*
U.S. History	9335	5469	857	13874	3754	22160	1012	45431	101892
	9.2%	*5.4%*	*0.8%*	*13.6%*	*3.7%*	*21.7%*	*1.0%*	*44.6%*	*100%*